XML::DOM メモ
XML::DOM を使う際の問題と対策

福原 知宏 (tomohi-f@cd5.so-net.ne.jp)
Kansai.pm 2**周年記念イベント**

問題と対策

- **問題**） XML::DOM **の出力が文字化けする**

```
$doc = new XML::DOM::Parser->parsefile(“test.xml”);
$title = $doc->getElementsByTagName(“title”)->…;
print “題名: $title¥n”;
```

XMLファイルを解析

文字化け

- **対策**） **文字コードを揃えましょう**

```
use Jcode;
print “結果¥n” . Jcode->new($title)->sjis;
```

文字コード変換モジュール

← OK

何故、文字化けするか？

- **文字コードが混在するため**
  - XML::DOM はUTF8 で結果を返す

$title = $doc->getElementsByTagName(“title”)->…;

UTF8**文字列**: Perl 5.6.1 **の内部では**UTF8 フラグ**がセットされている**

print “**題名**: $title¥n”; ← **異なる文字コードの混在！**

UTF8**文字列**

SJIS**文字列**: UTF8**フラグはセットされていない**

文字化けを防ぐ

- **文字コードを揃える**

use Jcode;

print “**結果**¥n” . Jcode->new($title)->sjis;

SJIS**文字列**　　SJIS**文字列**

文字コードの統一 → 文字化けの解消！

参考URL

- Perl のUnicode Support
  - http://homepage1.nifty.com/nomenclator/perl/unicode.htm
  - Perl 5.6.1 のUTF文字列の内部処理について説明している

UTF8 文字列の文字化け例題スクリプト

```
#!perl
#
#  元のコードは Shift JIS  で記述
# ActivePerl 5.6.1 build631
# Windows  環境で実行

use XML::DOM; # version 1.25
use Jcode;          # version 0.60

my $doc;
my $title;

# XML  文書の解析
my $parser = new XML::DOM::Parser;
$doc = $parser->parsefile("test.xml");
$title=$doc->getElementsByTagName("title")->item(0)->getChildNodes->item(0)->getN
odeValue(); # XML::DOM  の値は UTF8  で符号化されて返ってくる

#  文字化け: UTF8 文字列が表示される
print $title, "¥n";

#  文字化け: SJIS  と UTF8  文字列が混在している
print "結果: $title¥n";

#  文字化け: Jcode  に文字列を渡す前に文字コードが混在してしまっている
print Jcode->new("題名: $title¥n")->sjis;

#  今度は OK:  文字コードを SJIS  に揃えた
print "結果: " . Jcode->new($title)->sjis . "¥n";
```

XML 文書のサンプル(test.xml)

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<scenario version="1.0">
<header>
<title>新 Perl の国にようこそ</title>
<summary>本書は 1992 年に出版された「Perl の国へようこそ」を全面的に書き直したも
のです。</summary>
</header>
</scenario>
```